POWER AND COST CONSIDERATIONS FOR SMALL MAMMAL MONITORING

Eric Rexstad and Edward Debevec
Institute of Arctic Biology
University of Alaska Fairbanks
Fairbanks AK 99775-7000

April 1999

Contents

1.0 Introduction

Statistical power is defined as the probability of rejecting a null hypothesis when that hypothesis is false. When performing a statistical test, there are two possible outcomes; we either accept or reject the null hypothesis. Ideally we accept the null hypothesis when it is true and reject it when it is false. Unfortunately, this is not always the case and we are faced with the possibility of making two types of errors: rejecting the null hypothesis when it is true (Type I error) and accepting it when it is false (Type II error). The probability of making a Type I error is determined by the choice of α for the test and is thus independent of the experimental design. However, the probability of making a Type II error, also known as β, does depend on the experimental design. We do not customarily refer to β, but rather to power, which is equal to $1-\beta$. The following table may be useful in untangling the relationships of these terms.

		Decision	
		Accept Null	Reject Null
Null Hypothesis	True	Correct decision	Type I error (α)
	False	Type II error (β)	Correct decision ($1-\beta$ = Power)

It is clear that power is an important factor to be considered in a “successful” monitoring program that hopes to detect on-going trends or transient perturbations. If we are going to go to the trouble of implementing a monitoring program, we want to be confident that we can detect a change when it occurs. The difficulty arises when considering the type and magnitude of change to use for determining power because any discussion of power is only relevant for a clearly stated statistical test or comparison. Instead of calculating the power to detect any change, we can only calculate the power to detect a very specific change. We are therefore faced with the possibility that we could design a monitoring protocol based on attaining sufficient power to detect a particular type and magnitude of change, but then completely miss the change that actually occurs. On the one hand, we are asked to design a monitoring protocol that is all-encompassing and not tied to any given foreseeable change, while on the other hand we must consider very specific changes in order to talk about power. In this report, we will discuss some of the difficulties in determining the statistical power of a monitoring effort followed by an initial look at power relevant to small mammal monitoring in Denali LTEM.

2.0 Power and monitoring

By its very definition, power implies the use of a statistical test. Consider the following three examples that incorporate typical power considerations.

1. In a population study, we generate an annual estimate of r, the intrinsic rate of growth. The population is growing if $r > 1$ and declining if $r < 1$. We are primarily interested in knowing when r is less than 1 so we perform a statistical test with the following null hypothesis (H_0) and alternative hypothesis (H_A):

$$H_0: r \geq 1$$
$$H_A: r < 1$$

We talk about power as the probability of rejecting H_0 given that the true value of r is some value less than 1, a value that we have to specify. So for example, say we're interested in the probability of rejecting H_0 when the true value of r is 0.95 and we use $\alpha = 0.05$. The power we come up with is going to depend on how well we estimated r, i.e., the precision of our estimate as reflected in its standard error. Generally, the larger our sample size, the more precise our estimate and hence, the greater our power. We can alternatively specify our desired power and calculate the minimum sample size necessary to attain it. For a desired power of 80%, we determine the minimum sample size needed to reject H_0 80% of the time when the true r is 0.95. Two things are important to note here. First, there are no guarantees. Even with this sample size we will fail to reject H_0 20% of the time. Second, if the true r is between 0.95 and 1, then the power of this test is less than 80% despite that fact that the true r is less than 1. We had to discuss power with a very specific scenario in mind.

2. In the same population study, we also estimate an annual survival rate. We are interested in testing whether the survival rate in the first year of the study (S1) differs from that of the second year (S2). Our null and alternative hypotheses are as follows:

$$H_0: S_1 = S_2$$
$$H_A: S_1 \neq S_2$$

We talk about power in this case as the probability of rejecting H_0 given that the true value of S_1 is different than the true value of S_2. We have to go a step further and specify a difference between S_1 and S_2 that we want to detect, say 0.1. To couch the question again in terms of sample size, we determine the minimum sample size needed to reject H_0 80% of the time when the true difference between S_1 and S_2 is 0.1. The same cautionary notes apply here as before. The difference between S_1 and S_2 can be as much as 0.1 and we will fail to detect the difference 20% of the time, and there might be a true difference less than 0.1 and our power to detect the difference will be less than 80%.

3. In a monitoring context, we are often interested in trends over time. Say we now have 10 years of survival estimates and we want to test whether survival is decreasing

with time. A common approach is to regress survival on time and test for a slope less than 0. For a slope identified as β_1, our null and alternative hypotheses are as follows:

$$H_0: \beta_1 \geq 0$$
$$H_A: \beta_1 < 0$$

The discussion of power from example 1 above directly applies here, including the two cautionary notes. We will still miss this trend 20% of the time and it is possible to have a slope less than 1 that will be detected less than 80% of the time.

From a monitoring perspective, we want to be assured that we will detect a change when it occurs. Does power of 0.8 provide us with that assurance? We may want to set a higher level for power, say 0.95, that would instill in us more confidence that we would indeed detect the change. However, attaining that power may involve substantial increases in effort and cost to provide the requisite larger sample sizes. This will require cost-benefit analyses to determine whether this power level is feasible. And even at this level, someone may object to the fact that we could still be unable to detect the change 5% of the time. Basing a monitoring strategy on the probability of detecting a change that may not occur might not be the best way to proceed.

Talking about trends is particularly risky. In the third example above, we considered a linear trend over time where the attribute of interest steadily decreases over time. This is only one of many patterns that might occur over time. Consider instead a one-time perturbation to the system such as the Prince William Sound oil spill. We might expect to see a sudden drop in survival and then a gradual return to normal as the system recovers. Focusing on a linear trend analysis might cause us to completely miss the event. Or even if survival stays depressed for a long period of time, it may take several years before a test would show a statistically significant change. Instead, it could be enough to forgo the statistical test and simply demonstrate a relationship between the measured attribute and the perturbation. We could hope that the perturbation only affects some monitoring sites so that we could detect a sudden difference between sites following the insult. However, it is also possible that a perturbation will be system-wide so that between-plot comparisons reveal nothing when in fact, ecosystem-wide changes are occurring.

To summarize:

- Power deals with probabilities and hence the uncertainty of detecting a statistically significant change.
- Power requires an explicit test of hypothesis and does not have any relevance beyond that test and the scenario used to calculate it. The change that actually occurs may still have little chance of being detected.
- Depending on the variability of the system being measured, required sample sizes to attain desired levels of power may be unrealistic.
- Basing power on a single possible trend is too limiting in scope. When other patterns occur over time, there may be little possibility of detecting them.

3.0 Power and small mammal monitoring

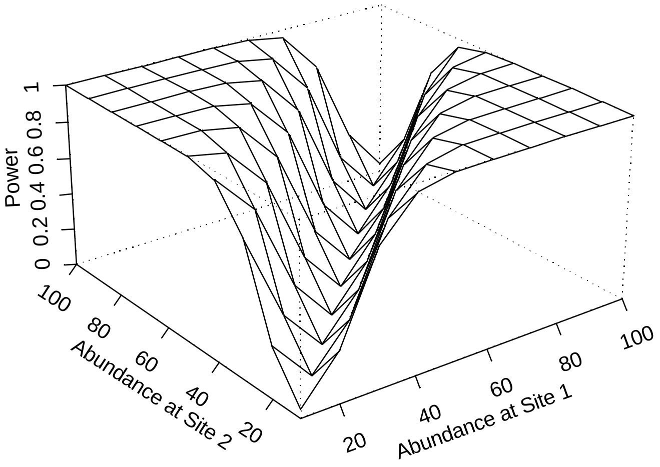
Applying power consideration to small mammal monitoring with the Denali LTEM program requires the consideration of population-level characteristics of the species of interest. Specifically, we must deal with the fact that they are irruptive species with abundance levels varying by as much as an order of magnitude from one year to the next, inducing highly inflated inter-annual variability. Questions regarding temporal differences in abundance become moot. The answer is an unequivocal "Yes, there are differences between years." For populations that maintain a constant abundance over time, this is a meaningful question that could lead to appropriate measures being taken when and if a difference is seen. For irruptive populations, such as small mammals in Denali, there is little point in asking the question because it is natural for their abundance to vary by these extreme amounts and the finding of differences between years does not compel any concern, much less any action.

Spatial patterns in population abundance or survival for an irruptive species could be of interest and worth investigating. Do population highs and lows tend to occur at the same time for different locations within Denali or are these population attributes independent across space? Do population attributes follow a discernable gradient or are they fairly consistent across space? These are questions that make sense when dealing with the sort of variability we see with small mammals in Denali and will allow us a context for a discussion of power.

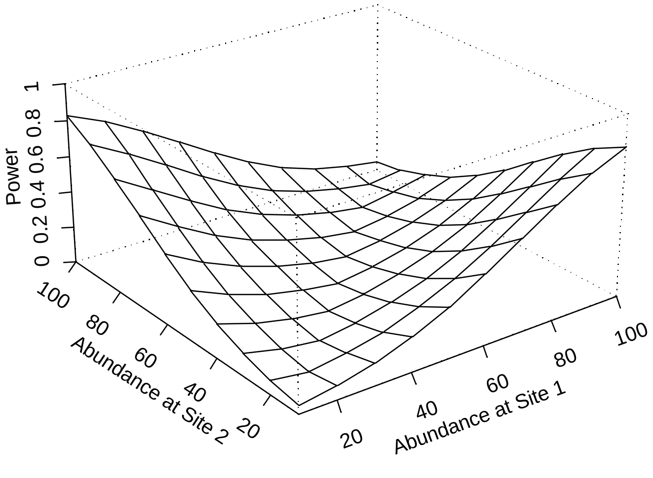
The best approach to this sort of analysis is to use simulations to duplicate the entire sampling process. Current sampling protocol involves five primary sampling events per plot during the summer with each primary sampling event consisting of 12 secondary sampling events (3 per day × 4 days). There are other factors that need to be considered as well, most important of which are the capture probabilities. We recently performed a suite of simulations to examine the effect of varying the number of secondary sampling events on the quality of the resulting abundance estimates. Capture probabilities were simulated using models M_0 (all capture/recapture probabilities equal for all individuals and constant over time), M_b (probability of first capture is different from probability of recapture for all individuals), and M_h (capture/recapture probabilities differ by individual, but constant over time). We calculated abundance estimates for 2 to 18 secondary sampling events and several scenarios under each model. In general, estimates had greater variability and more bias at low numbers of secondary sampling events (2 to 5), but stabilized with more secondary events. Performance greatly depended on the capture probability or range of probabilities used. For greater probabilities, as few as 8 secondary sampling events appeared to be sufficient. However, this did not hold up at a minimum capture probability of 0.1. Our conclusion from this analysis was that 12 was the appropriate number of secondary events and would be adequate for all capture probabilities down to our minimum of 0.1. All further analyses will be done using 12 secondary sampling events and the worst-case capture probability model (M_h with capture probabilities from 0.1 to 0.3). We will focus on abundance as the attribute of interest for these simulations and examine our power to detect changes and trends in abundance across time and/or space.

3.1 Detecting differences between two locations - Methods

We first consider small mammal sampling for two locations and the power to detect a difference in abundance between them. To date we have monitored small mammal populations at the east end of the park road (near headquarters) and at the west end (near Wonder Lake) and we would like to test for differences between them. We performed a series of simulations for two populations, each with a range of abundances from 10 to 100. Capture probabilities were generated and abundance estimates calculated. This was repeated for 500 reps with the proportion of reps resulting in a significant difference found between populations taken as an estimate of power. Results will be used to generate a power surface showing the power to detect differences in survival.



Ideally, we get a power surface such as this where we have pretty good power to detect even slight differences in survival. The diagonal trough in this plot indicates no difference between survival at the two locations. This is low (the value should be equal to α) because we do not want to detect a difference here. As we move off the diagonal, the survivals are different and we hope to detect that difference. The steeper the sides of the trough, the smaller the difference we are able to detect with reasonable certainty.



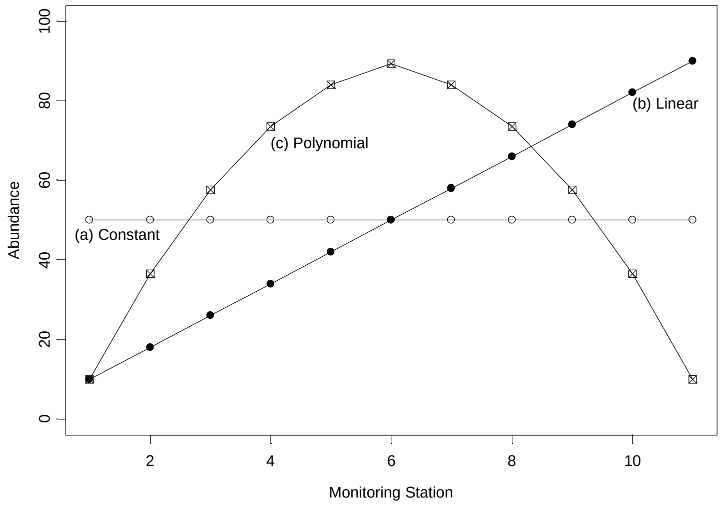
Alternatively, we could end up with a power surface such as this where we are only able to detect extreme differences in survival with any certainty. This sampling design would stand little chance of detecting small or moderate differences in survival. In this situation, there may be little that can be done to improve the power to detect differences given the high variability of attributes of irruptive species and the need to keep the length of a primary sampling period short to maintain a closed population.

With good power to detect differences between locations, we stand a good chance of detecting localized impacts that would only affect one population while leaving the other as a control. However, a system-wide impact could affect both populations equally, making it questionable whether we could detect it.

3.2 Detecting trends among multiple locations - Methods

Although a trend analysis over time would not be very valuable as explained above, a trend analysis along a transect might be very interesting. We use the sampling framework we discussed previously with regards to scaling-up options in Denali LTEM where we proposed the selection of 12 sampling locations along the road corridor. The road corridor essentially constitutes a transect along which we can look for spatial gradients or other patterns of population attributes. For these simulations, we considered 11 sampling locations instead of the proposed 12 to facilitate defining and categorizing trends along the transect.

We again used simulation methods to consider our power to detect various patterns in abundance along this transect. We induced a structure to the true abundance at each of the 11 sampling locations and generated capture histories that were used to estimate abundance. Then we used regression techniques to test whether a trend can be found. Induced patterns were (a) constant abundance and (b) linearly increasing abundance. A third pattern, (c) abundance greater at the middle of the transect and lower at both ends, will be included later. The basic forms of the induced trends are shown here.



The regression analysis tests for a linear and a quadratic trend. Other trends come to mind, such as constant abundance over half the transect and increasing/decreasing over the second half, but they will not be considered here. The patterns mentioned should be sufficient to address the issue. Each pattern will be used with several ranges of abundance as indicated in the following table.

	End 2 (b) or Middle (c)									
End 1	10	20	30	40	50	60	70	80	90	100
10	a	b, c	b, c	b, c	b, c	b, c	b, c	b, c	b, c	b, c
20		a	b, c	b, c	b, c	b, c	b, c	b, c	b, c	b, c
30			a	b, c	b, c	b, c	b, c	b, c	b, c	b, c
40				a	b, c	b, c	b, c	b, c	b, c	b, c
50					a	b, c	b, c	b, c	b, c	b, c
60						a	b, c	b, c	b, c	b, c
70							a	b, c	b, c	b, c
80								a	b, c	b, c
90									a	b, c
100										a

Induced abundances along the transect were interpolated from the values given in the table. Constant abundance (i.e., no trend) used abundances from 10 to 100. Each linear trend (b) used 11 abundances evenly spaced and increasing from the value for End 1 to that of End 2. Quadratic trends (c) will use six abundances evenly spaced and increasing from the value for End 1 to that of Middle, with the remaining five abundances mirroring the first five. Results from (b) and (c) series simulations are symmetrical so the lower half of the table will mirror the upper half results. Power surface plots for linear trends were created as described above.

3.3 Detecting differences between two locations - Results

Abundances were varied from 10 to 100 in increments of 1. Capture histories were generated for each abundance level using the M(h) model with capture probabilities from 0.1 to 0.3. Abundance estimates with standard errors were computed using CAPTURE. This was replicated 500 times. A simple Z-test was used to test for differences (α = 0.05). For tests where abundance differed, all 500 replicates from each abundance level were used. For tests where abundance did not differ, the 500 replicates were divided into two groups, resulting in a diminished sample size of 250. Results are shown in Table 1 for abundances that are multiples of 10.

Table 1: Simulation results (reduced set) for detecting differences in abundance between two locations. Power values given are the probability of detecting a significant difference. Shaded cells on the diagonal represent equal abundances at both locations and should have power equal to the α used (0.05).

	Location 1										
	N	10	20	30	40	50	60	70	80	90	100
Location 2	10	0.016	0.716	0.908	0.966	0.974	0.980	0.986	0.990	0.986	0.992
	20	0.716	0.004	0.494	0.920	0.950	0.964	0.976	0.986	0.986	0.988
	30	0.908	0.494	0.024	0.400	0.838	0.926	0.950	0.970	0.978	0.978
	40	0.966	0.920	0.400	0.004	0.334	0.828	0.930	0.956	0.966	0.964
	50	0.974	0.950	0.838	0.334	0.012	0.210	0.756	0.874	0.928	0.940
	60	0.980	0.964	0.926	0.828	0.210	0.000	0.172	0.712	0.880	0.928
	70	0.986	0.976	0.950	0.930	0.756	0.172	0.004	0.174	0.674	0.866
	80	0.990	0.986	0.970	0.956	0.874	0.712	0.174	0.004	0.126	0.578
	90	0.986	0.986	0.978	0.966	0.928	0.880	0.674	0.126	0.016	0.104
	100	0.992	0.988	0.978	0.964	0.940	0.928	0.866	0.578	0.104	0.004

Full results are plotted on the following page. Figure 1 provides a three-dimensional view of the power curve, while Figure 2 shows the same data as a contour plot that is more easily read. For example, with an abundance of 60 animals at one location and 80 animals at another, our power to detect a difference is approximately 0.7. Power to detect a given difference between locations appears to decrease as the absolute abundances increase (i.e., the diagonal trough in the power surface widens as abundance increases).

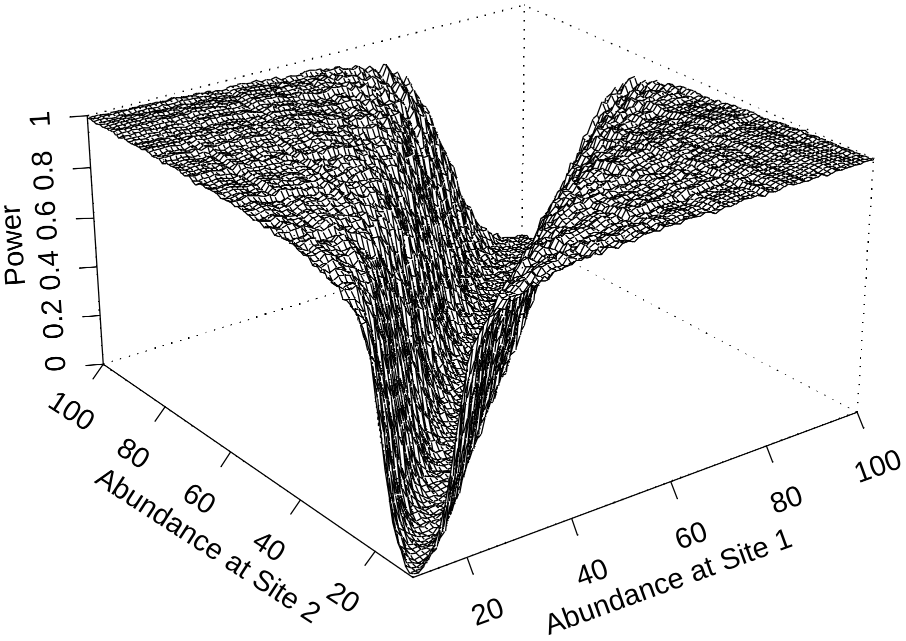


Figure 1: Power surface for detecting differences in abundance between two locations.

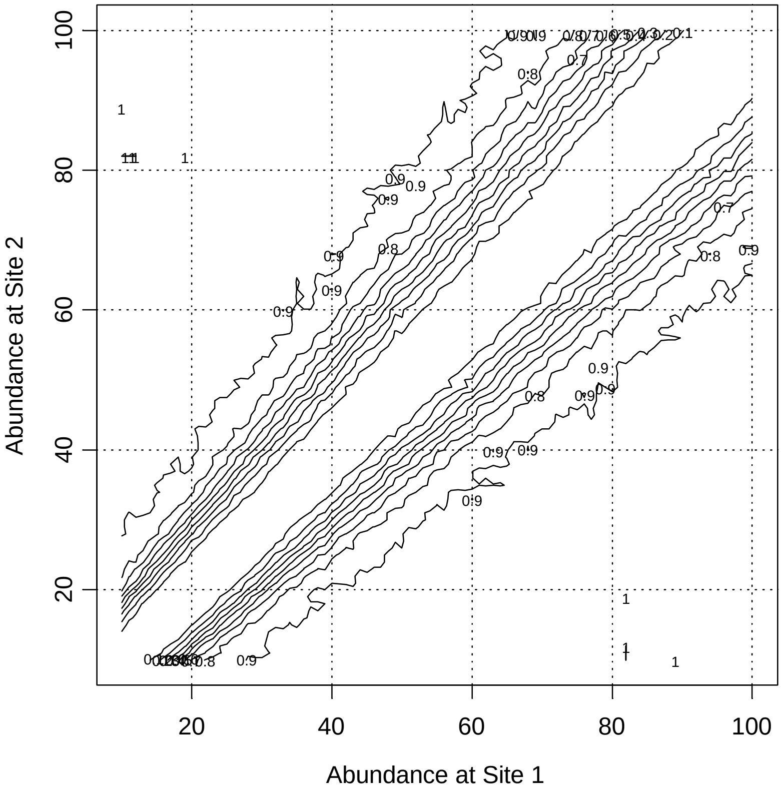


Figure 2: Power contour plot for detecting differences in abundance between two locations.

3.4 Detecting trends among multiple locations - Results

Abundances were varied between 10 and 100 in increments of 1. Each simulation is categorized by two values: starting abundance and slope. The starting abundance is the abundance at the first locations along the transect, which is also designated as the "low" end of the trend. The slope is the change in abundance from one location to the next. So with a starting abundance of 10 and a slope of 1, abundances would range from 10 at one end to 20 at the other. A slope of 2 would indicate abundances from 10 to 30, etc. Not all combinations of starting abundance and slope are possible. For example, to start at 60 with a slope of 5 would require abundances greater than 100. Only those linear trends with populations between 10 and 100 were run.

For each simulation, a set of capture histories was generated for each abundance level using the M(h) model with capture probabilities from 0.1 to 0.3. Abundance estimates with standard errors were computed using CAPTURE. This was replicated 500 times. The resulting 11 abundance estimates from each simulation were then analyzed with simple linear regression and tested for a slope significantly different from zero ($\alpha = 0.05$). If a trend had locations with the same abundance, then bootstrapping with replacement was used to generate a new series of abundance estimates. Results are shown in Table 2, again only for those starting abundances that are multiples of 10.

Table 1: Simulation results (reduced set) for detecting a linear trend in abundance along a transect. Power values given are the probability of detecting a significant trend. Cells in the first column represent no trend in abundance should have power equal to the α used (0.05).

		Slope of Linear Trend									
		0	1	2	3	4	5	6	7	8	9
Starting Abundance	10	0.030	0.840	0.934	0.990	0.998	1.000	1.000	1	1	1
	20	0.036	0.866	0.948	0.984	0.990	0.998	1.000	1	1	-
	30	0.036	0.822	0.938	0.972	0.988	0.996	1.000	1	-	-
	40	0.054	0.780	0.940	0.964	0.984	0.992	0.998	-	-	-
	50	0.030	0.732	0.948	0.962	0.982	0.996	-	-	-	-
	60	0.060	0.654	0.944	0.972	0.980	-	-	-	-	-
	70	0.068	0.598	0.932	0.960	-	-	-	-	-	-
	80	0.050	0.568	0.946	-	-	-	-	-	-	-
	90	0.048	0.544	-	-	-	-	-	-	-	-
	100	0.056	-	-	-	-	-	-	-	-	-

Full results are plotted on the following page. Figure 3 provides a three-dimensional view of the power curve, while Figure 4 shows the same data as a contour plot that is more easily read. For example, with abundances less than 50 animals and a slope of only 1, our power to detect a linear trend is approximately 0.8. Power to detect a trend appears to decrease as the absolute abundances increase above 50.

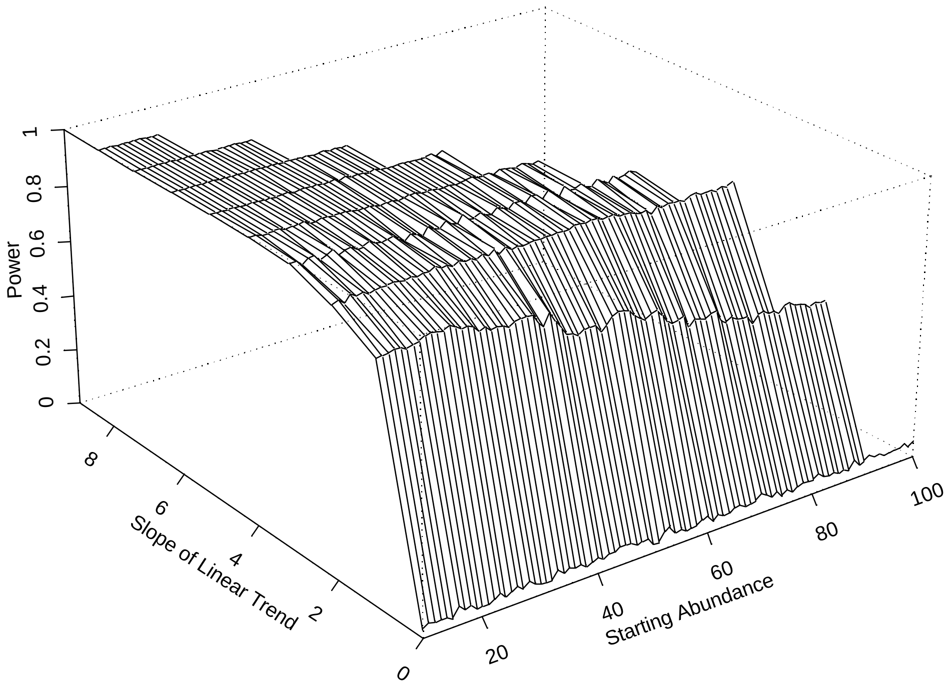


Figure 3: Power surface for detecting a linear trend in abundance along a transect.

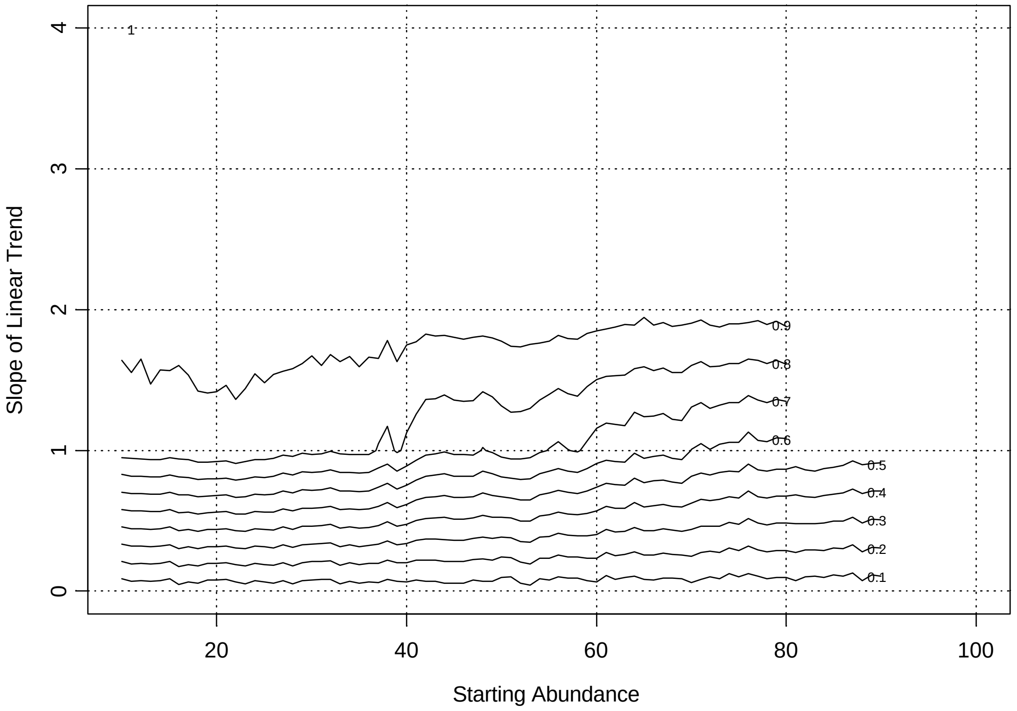


Figure 4: Power contour plot for detecting a linear trend in abundance along a transect.

4.0 Conclusion

Power considerations begin with the individual abundance estimates computed for each plot and sampling occasion. Adjustments can be made to the number of secondary sampling events carried out during a single primary sampling event. We have determined that our current strategy of 12 secondary events carried out over 4 days is sufficient to obtain good precision in our estimates.

Power to detect a difference between two locations appears reasonable. Using the common value for power of 0.8, we can generalize what magnitude of difference we can detect. Table 3 summarizes the detectable differences for a range of abundances. For example, for a location with an abundance of 60 animals, we have at least 80% power to detect a difference from locations with less than 40 or greater than 82 animals.

Table 3: Detectable differences for select abundance levels with a minimum power of 80%. Values determined from Figure 2. Values identified with an asterisk (*) were extrapolated.

	Detectable Difference	
Abundance	Below	Above
10	< 0*	> 22
20	< 8*	> 34
30	< 16	> 46
40	< 24	> 58
50	< 32	> 70
60	< 40	> 82
70	< 48	> 94
80	< 56	> 106*
90	< 64	> 118*
100	< 72	> 130*

Power to detect a linear trend along a transect also appears reasonable. A trend with a slope of 1 indicates a difference in abundance of 10 animals between one end of the transect and the other. Power is greater than 80% for a linear slope of 1 and abundances less than 50. Power is greater than 90% to detect a linear trend with a slope of 2 for all abundances used in these simulations. A slope of 2 indicates a difference in abundance of 20 between the ends of the transect.

These exercises give us a general sense of whether we will be able to detect some changes that may occur in the future. We may not be able to guarantee that we will detect every possible change with Denali LTEM, but we can have confidence that the data we do collect will allow us to detect some acceptable level of change.